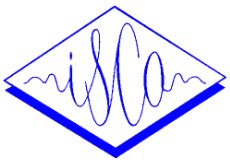

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A NEW JAPANESE TTS SYSTEM BASED ON SPEECH-PROSODY DATABASE AND SPEECH MODIFICATION

Mitsuaki ISOGAI, Kimihito TANAKA, Satoshi TAKANO,
Hideyuki MIZUNO, Masanobu ABE, and Sin'ya NAKAJIMA
NTT Cyber Space Labs.

1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
isogai@nttspch.hil.ntt.co.jp

ABSTRACT

This paper describes a new Japanese text-to-speech (TTS) system that can produce highly natural and intelligible synthetic speech. The good performance of the new TTS system derives from three new sophisticated approaches as follows; (1)A new prosody control algorithm that uses prosody data extracted from a natural speech database and a duration control algorithm based on statistical estimation. (2)A new type of synthesis unit that consists of a consonant with following vowel chain. The unit suppresses unnatural sounds and acoustic discontinuities at concatenation points by preparing synthesis units with various lengths and various F0 contours. (3)A new speech modification algorithm with harmonics reconstruction. To evaluate the new modules and the total performance of the new TTS system, listening tests are carried out. The results confirm that the new modules work together effectively, and that the new TTS system can produce high quality synthesized speech.

1. INTRODUCTION

We developed a TTS system based on waveform concatenation in 1995[1]. That system is being used in many applications such as a telephone banking service, a telephone information service, multimedia systems, and character voices in computer games. The synthesized speech produced by that system has highly intelligibility, but its naturalness is rather weak. To extend the application area of TTS systems, we must improve the naturalness of synthesized speech. For human-like synthesized speech, improvements are required in these areas of prosody control, synthesis units, and speech modification.

Many improved prosody control methods have been recently introduced, but their output is still monotonous. To produce more fluent speech, corpus-based statistical estimation models such as the HMM-based F0 generation model[2] and the two-stage F0 control model[3] have been proposed. We also examined the use of prosodic data extracted from a natural speech database[4]. Research into synthesis units was advanced with the proposal of non-uniform units; they permit some current systems to offer more natural synthetic speech[5][6]. We also proposed multi-form units[7] to reduce acoustic discontinuities at concatenation points. Even if the synthesis unit database is optimally designed, the number of speech variations is insufficient, and the difficulty of modifying the speech to generate natural prosody degrades speech quality. Although the TD-PSOLA[8] algorithm is the most popular synthesis algorithm in current TTS systems, it still has problems such as a relatively narrow range of modification wherein naturalness is retained. To overcome this problem, some new approaches such as STRAIGHT[9], HNM[10] and our HARP [11] algorithm were proposed.

Our prior proposals introduced a new method of prosody control, synthesis units, and speech modification. This paper integrates these three techniques to develop a new Japanese TTS system. In order to clarify the improvements achieved, we examine the performance of each module, the effect of module combination, and the total performance of the TTS system. The following section overviews our new TTS system.

2. SYSTEM OVERVIEW

Figure 1 overviews the new TTS system. The TTS system has four modules: text analysis, prosody control, synthesis unit selection, and speech synthesis. In the text analysis module, the input Japanese text, which consists of Kanji and Kana sequences, is transformed into phonetic symbols. These phonetic symbols are divided into Japanese minor phrases. In addition, each minor phrase is assigned accent type and boundary type. The prosody control module sets the F0 contours and duration of each phoneme. In the unit selection module, the most suitable stored speech synthesis unit is selected. The speech

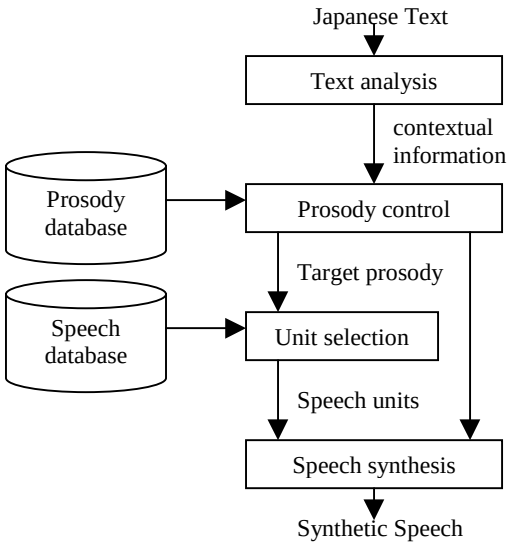

Figure 1: System overview

synthesis module modifies the F0 contours and duration of the speech segments to match those of the target; they are concatenated to synthesize the target speech. The following section details the prosody control module. Section 4 explains the unit selection module. Section 5 examines the speech synthesis module.

3.PROSODY CONTROL MODULE

3.1 Prosody database

We previously reported that F0 contour is greatly influenced by contextual information[4]. We defined the contextual information as accent type, the number of syllables, the phonetic sequence of the minor phrase, and so on. The contextual information can be obtained from the text analysis module. Many factors related to contextual information affect F0 contour. Therefore, according to an analysis of the influence of contextual information factors on F0 contour[3][4], we determined the major factors shaping F0 contour. Using these factors, we designed a prosody database that covers most F0 contour types in Japanese minor phrases. The factors are shown in Table 1. The prosody database in the new TTS system has 1440 F0 contour types.

3.2 F0 contour selection and location

Each F0 contour is selected according to the contextual information of the target phrase. The most suitable F0 contour, the one whose contextual information most closely matches the target, is selected from the prosody database. If the target phrase has more than 9 syllables, the F0 contour, which has 8 syllables, is extended to match the target phrase. The dynamic range and height of the target phrase are determined by F0 generation rules[1].

Table 1: Factors and categories of the prosody database

factors		categories
the number of syllables		from 1 through 8
accent position	preceding phrase	none, exist
	current phrase	none, from 1st syllable to 7th syllable
boundary type	preceding	tight or loose connection, preceded by pause
	following	followed by pause, end of a sentence
location of minor phrase		top, second, over third

Table 2: Multiple correlation coefficients and RMS errors of duration estimation

current phoneme	multiple correlation coefficients	RMS error (msec)
short vowel	0.705	11.727
long vowel	0.557	18.620
consonant	0.559	10.567
nasal	0.694	14.324

3.3 Duration control

The duration of each phoneme in the target text is determined by statistical estimation theory type I [12]. A speech database that contains 503 phonetically balanced sentences[13] was used in this estimation. 5 factors were adopted for estimation: preceding and following phonemes, the number of syllables in current minor phrase, the location of current phoneme in current minor phrase, and the location of current minor phrase in the gross phrase. The estimations were carried out for four cases: current phoneme is a short vowel, long vowel, consonant, or nasal. The multiple correlation coefficients and RMS errors of the estimation are shown in Table 2.

4. UNIT SELECTION MODULE

4.1 Multi-form unit

Japanese syllables are either CV(Consonant-Vowel sequence) or V, and it is well known that transitions from C-to-V, or from V-to-V are very important in terms of auditory perception. Therefore, the quality of synthetic speech is strongly degraded if speech segments are concatenated at C-to-V or V-to-V transients because of the acoustic discontinuities so formed. Multi-form synthesis units were proposed with the intention of reducing these acoustic discontinuities[7]. Each unit is a consonant with following vowel chain. Here, the vowel chain can be any sequence of vowels, semivowels, or syllabic nasals. This unit is identified by the symbol sequence C(V)k, where k denotes the number of vowels in the vowel chain. Moreover, to realize smooth concatenation, we consider both the previous and succeeding phoneme environments of the C(V)k units. To minimize the decrease in speech quality caused by strong F0 modification, one C(V)k unit can take various F0 contours. We collected 50,000 of the most common units covering 75% of Japanese texts. We also prepared a simple diphone unit set that can cover all Japanese phoneme sequences, because the 50,000 C(V)k units do not cover all phoneme sequences in Japanese.

4.2 Unit selection

To use multi-form units for TTS, unit selection is performed using the speech database. First, we attempt to locate suitable C(V)k units. Every C(V)k unit appearing the target phoneme sequence is searched for in the speech database. The most suitable speech segment, the one whose F0 contour most closely matches the target, is selected from the candidates. If C(V)k search fails, diphone units are used for synthesis.

5.SPEECH SYNTHESIS MODULE

5.1 Unit concatenation

The speech synthesis module consists of two parts: the segment concatenation part and the F0 modification part. The former uses one of the two below methods, where selection depends on the segments being concatenated.

(case 1) Two C(V)k units are concatenated at the phonome boundaries.
(case 2) A C(V)k unit and a diphone unit, or two diphone units are concatenated within the phoneme. The concatenation point in each phoneme changes with the features of the phoneme.

5.2 Hybrid synthesis

We proposed a new speech modification algorithm (HARP) based on a vocoder framework[11]. The novel point of HARP is its ability to reconstruct spectrum harmonics according to the target F0. HARP offers better quality than TD-PSOLA[8] if the modification is large. However, in case of narrow F0 modification, HARP and TD-PSOLA yield similar performance. HARP is more expensive than TD-PSOLA in terms of computational cost. Therefore, we introduce here a hybrid synthesis system in which HARP and our conventional PSOLA-like algorithm are applied for strong and slight F0 modification, respectively.

6. EVALUATIONS

This section details three evaluations. To evaluate the performance of each module, preference tests were carried out. To evaluate the performance of the combined modules, intelligibility and opinion tests were carried out.

6.1 Conditions

Experimental conditions are shown in Table 3. In hybrid synthesis, our conventional PSOLA-like algorithm was applied when C(V)k unit search was successful and the F0 modification range was 0.8(downward modification) to 1.4(upward modification). The HARP algorithm was applied in all other cases.

6.1.1 Preference tests

Prosody control algorithm
Preference tests were carried out to evaluate the naturalness of proposed prosody control algorithm which uses prosody database. Sentences were synthesized using the proposed prosody control method and the conventional point-pitch model in which the accent component of a minor phrase is added to a gross phrase component and phoneme duration is determined by previous and succeeding phoneme environment. All of these speeches were synthesized using multi-form units.

Multi-form units
To evaluate the advantages of multi-form units, preference tests were carried out to evaluate the naturalness of the synthesized speech. For fair comparison, the conventional synthesis and prosody algorithms used two types of units: multi-form units and triphone units. In addition, to examine the effect of sampling rate difference, we carried out this test using conventional units with sampling rates of 11.025KHz and 22.05KHz.

6.1.2 Word Intelligibility

A word intelligibility test using Japanese family names was carried out. Half the names were familiar, and the remainder were unfamiliar. Most names consisted of 3, 4 or 5 syllables. The names were synthesized in the 5 ways shown in Table 4.

6.1.3 Opinion test

Opinion tests were carried out to evaluate the total performance of the new TTS system. The sentences were synthesized in the 7 different ways shown in Table 5.

Table 3: Experimental conditions

New TTS System	
Number of C(V)k units Number of diphone units	50,000 10,000
Sampling rate	22.05 KHz
Prosody control	prosody database or point-pitch
Speech synthesis	HARP or PSOLA-like or hybrid

Conventional TTS System	
Number of triphone units	6,000
Sampling rate	11.025 KHz or 22.05 KHz
Prosody control	point-pitch
Speech synthesis	PSOLA-like

Subjects	10 females
Sentences (preference and opinion tests)	5 common Japanese sentences
Words (intelligibility test)	50 Japanese family names

Figure 2: Preference test results of prosody module.

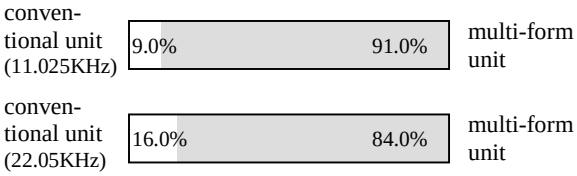

Figure 3: Preference test results of multi-form unit.

6.2 Results and discussion

The results of the preference test of prosody control algorithms are shown in Figure2. 76% of the subjects judged that the proposed method generated more natural prosody than the conventional method.

Figure3 shows the results of the preference test of multi-form units. 91% (84%) of the words were judged by the subjects as indicating that the synthetic speech based on multi-form units was more natural than the speech based on conventional units at the sampling rate of 11.025KHz (22.05KHz). This difference in preferences suggests that raising the sampling rate improves the naturalness of conventional units. Even if the sampling rate is raised, the multi-form units offer better naturalness than the conventional units.

Table 4 shows the intelligibility scores recorded in the word intelligibility test. (a)The conventional TTS system offered superb intelligibility because the triphone units were designed to achieve highly intelligibility. System (e) also offered excellent intelligibility. The difference of scores between (b) and (d) was due to the difference in speech modification range; the

Table 4: Intelligibility scores

(a) Conventional TTS system		99.2 %
New TTS system	(b) point-pitch and PSOLA -like	98.7 %
	(c) point-pitch and HARP	96.3 %
	(d) prosody database and PSOLA-like	97.1 %
	(e) prosody database and HARP	99.0 %

Table 5: Opinion scores

(a) Conventional TTS system		1.89
New TTS system	(b) point-pitch and PSOLA-like	3.00
	(c) point-pitch and HARP	2.57
	(d) point-pitch and hybrid	3.20
	(e) prosody database and PSOLA-like	3.03
	(f) prosody database and HARP	2.89
	(g) prosody database and hybrid	3.29

prosody database generates more pronounced intonation than the conventional system, so the speech quality of (d) decreased with large F0 modification. The difference in scores between (d) and (e) indicates that HARP synthesis works well. While the prosody database with the PSOLA-like algorithm didn't increase intelligibility, the prosody database with HARP did. This result confirms that the combination of these modules realized the advantage of each module. As for (c), we discuss the reason for its lower score below.

The result of the opinion test is shown in Table 5. It indicates that (a)the conventional TTS system yielded the lowest score. The preference test also confirmed that the conventional TTS system tends to lacks naturalness. As regards the prosody control method, even if the same speech synthesis module is used, the prosody database method is more natural than the conventional point-pitch method. As regards the speech synthesis module, the hybrid method was more natural than the PSOLA-like method indicating that the hybrid module works well. However, the HARP method was less effective than the PSOLA-like method because several phonemes in the test set were not well handled by the HARP method. This problem is overcome with the hybrid method. System (g) yielded the most natural speech confirming that the combination of our new modules works well.

7. CONCLUSION

A new Japanese TTS system based on a speech-prosody database and speech modification was described. The new TTS system has three parts: F0 control algorithm that uses a natural prosody database, multi-form synthesis units, and a speech modification algorithm with harmonics reconstruction. We carried out evaluations to test the effectiveness of each module and the total performance of the new TTS system. The results confirm that the new TTS system can produce synthetic speech that has high intelligibility and naturalness, and that the combination of the modules is very effective. The combination of these three parts not only improves the speech quality of TTS systems, but also realizes the generation of emotional speech, control voice quality, and change speaker identify, all of which are expected to become essential in the next generation of speech synthesizers.

8. ACKNOWLEDGEMENT

We are grateful to the members of the Media Processing Project for their helpful discussions. We also thank Mr.Yamamori, the executive manager, for his continuous support of this work.

9. REFERENCES

1. K.Hakoda, T.hirokawa, H.Tsukada, Y.Yoshida, H.Mizuno,"Japanese text-to-speech software based on waveform concatenation method," Proc.AVIOS'95, pp.65-72, 1995
2. T.Fukada, Y.Komori, T.Aso, Y.Ohora, "A study on pitch pattern generation using HMM-based statistical information," Proc.ICSLP'94, pp.723-726, 1994
3. M.Abe, H.Sato, "Two-stage F_0 control model using syllable based F_0 units," Proc.ICASSP'92, pp.118-121, 1992
4. M.Isogai, H.Mizuno, "A new F0 contour control method based on vector representation of F0 contour," Proc. Eurospeech'99, pp.727-730 ,1999
5. Y.Sagisaka, "Speech synthesis by rule using optimal selection of non-uniform systhesis," Proc.ICASSP'88, pp.679-682, 1988
6. A.Black, N.Campbell, "Optimizing selection of units from speech database for concatenative synthesis," Proc. Eurospeech'95, pp.581-584, 1995
7. K.Tanaka, H.Mizuno, M.Abe, S.Nakajima, "A Japanese text-to-speech system based on multi-form units with consideration of frequency distributionin Japanese," Proc.Eurospeech'99, pp.839-843, 1999
8. E.Moulines, F.Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, vol.9, no.5/6, pp.453-467, 1990
9. H.Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum : Vocoder revisited," Proc.ICASSP'97, pp.1303-1306, 1997
10. Y.Stylianou, T.Dutoit, J.Schroeter, "Diphone concatenation using a harmonic plus noise model of speech," Proc.Eurospeech'97, pp.613-616, 1997
11. S.Takano, M.Abe, "A new F0 modification algorithm by manipulating harmonics of magnitude spectrum," Proc. EUROSPEECH'99, pp.1875-1878, 1999
12. C.Hayashi, "On the quantification of qualitative data from the mathematicostatistical point of view", Ann. Inst. Statist., 1950
13. M.Abe, Y.Sagisaka, H.Kuwabara, "Fundamental frequency database with linguistic phonetic information," JASA, Vol.86 Supll.1 O8 pp.S36, 1989